

Empirical Processes, Typical Sequences and Coordinated Actions in Standard Borel Spaces

Maxim Raginsky, *Member, IEEE*

Abstract—This paper proposes a new notion of typical sequences on a wide class of abstract alphabets (so-called standard Borel spaces), which is based on approximations of memoryless sources by empirical distributions uniformly over a class of measurable “test functions.” In the finite-alphabet case, we can take all uniformly bounded functions and recover the usual notion of strong typicality (or typicality under the total variation distance). For a general alphabet, however, this function class turns out to be too large, and must be restricted. With this in mind, we define typicality with respect to any Glivenko–Cantelli function class (i.e., a function class that admits a Uniform Law of Large Numbers) and demonstrate its power by giving simple derivations of the fundamental limits on the achievable rates in several source coding scenarios, in which the relevant operational criteria pertain to reproducing empirical averages of a general-alphabet stationary memoryless source with respect to a suitable function class.

Index Terms—Coordination via communication, empirical processes, Glivenko–Cantelli classes, rate distortion, source coding, standard Borel spaces, typical sequences, uniform laws of large numbers.

I. INTRODUCTION

The notion of *typical sequence* has been central to information theory since Shannon’s original paper [1]. For finite alphabets, it leads to simple and intuitive proofs of achievability in a wide variety of source and channel coding settings, including multiterminal scenarios [2]. Another appealing aspect of typical sequences is that they provide a language for *approximation* of information sources in total variation distance using finite communication resources. Recent work of Cuff et al. [3] on coordination via communication serves as a particularly striking example of the power of this language.

For abstract alphabets, however, most of this power is lost; while such results as the asymptotic equipartition property carry over [4], in most other situations, particularly involving lossy codes, one has to resort to ergodic theory [5] or large deviations theory [6]. Direct approximation of abstract memoryless sources in total variation using empirical distributions is, in general, impossible (cf. Sec. IV for details). However, it is precisely this direct approximation that renders typicality-based proofs of achievability so transparent.

The present paper makes two contributions. First, we propose a way to revise the notion of typicality for *general* alphabets (more specifically, standard Borel spaces [7], [8]),

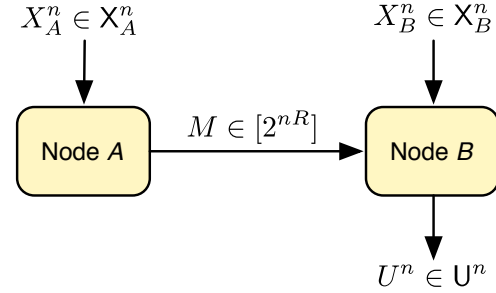


Fig. 1. Empirical coordination of actions in a two-node network. Node A (resp., B) observes a random n -tuple X_A^n (resp., X_B^n), where $(X_{A,1}, X_{B,1}), \dots, (X_{A,n}, X_{B,n})$ are i.i.d. pairs of correlated random variables. A message is sent from Node A to Node B at rate R to specify the n -tuple U^n .

allowing for similarly transparent achievability arguments. When two probability measures are close in total variation, the corresponding expectations of *any* bounded measurable function are also close. For general alphabets, when one of the measures is discrete, this is too much to ask. Instead, we advocate an approach based on suitably *restricting* the class of functions on which we would like to match statistical expectations with sample (empirical) averages. Provided the Law of Large Numbers holds *uniformly* over the restricted function class, we can speak of typical sequences *with respect to this class* and develop typicality-based achievability arguments in close parallel to the finite-alphabet case. The central object of study is the *empirical process* [9]–[11] indexed by the function class, which gives information on the deviation of empirical means from statistical means for a given realization of the source under consideration, and the total variation distance is replaced by the supremum norm of this empirical process.

The second contribution consists of applying our new notion of typicality to several source coding problems which, following the terminology of [3], can be thought of as “empirical coordination” of actions in a two-node network. Roughly speaking, the objective is to use communication resources in order to reproduce (or approximate) the empirical distribution of a given source sequence, rather than the sequence itself, with or without side information. This coordination viewpoint suggests a new operational framework suitable for problems involving distributed learning, control, and sensing.

A. Preview of the results

Consider the two-node network shown in Figure 1. There is an alphabet X_A associated with Node A , and two alphabets,

A preliminary version of this work was presented at the IEEE International Symposium on Information Theory, Austin, TX, July 2010.

M. Raginsky is with the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA. E-mail: m.raginsky@duke.edu.

X_B and U , associated with Node B . Initially, Node A (resp., Node B) observes a random n -tuple $X_A^n \in \mathcal{X}_A^n$ (resp., $X_B^n \in \mathcal{X}_B^n$), where the pairs $(X_{A,1}, X_{B,1}), \dots, (X_{A,n}, X_{B,n})$ are i.i.d. draws from some specified probability law $P_{X_A X_B}$ on $\mathcal{X}_A \times \mathcal{X}_B$. We also have a target conditional probability law $P_{U|X_A X_B}$ on U given X_A and X_B . Node A , given its knowledge of X_A^n , $P_{X_A X_B}$, and $P_{U|X_A X_B}$, communicates some information M to Node B at rate R . The latter receives M and, using its knowledge of X_B^n , $P_{X_A X_B}$, and $P_{U|X_A X_B}$, generates an n -tuple $U^n \in U^n$.

Now imagine that there is an external observer with access to X_\bullet^n (where \bullet is either A or B) and U^n , who also knows $P_{X_A X_B}$ and $P_{U|X_A X_B}$. This observer has a collection \mathcal{F} of “test functions” $f : \mathcal{X}_\bullet \times U \rightarrow [-1, 1]$ and can compute the *empirical expectation* (or sample average) $n^{-1} \sum_{i=1}^n f(X_{\bullet,i}, U_i)$ and the “true” expectation $\mathbb{E}f(X_\bullet, U)$ w.r.t. the joint law $P_{X_A X_B U} = P_{X_A X_B} \otimes P_{U|X_A X_B}$ for any $f \in \mathcal{F}$. We assume that Nodes A and B know \mathcal{F} , but do not know which $f \in \mathcal{F}$ the observer will pick. The objective is then to minimize the expected worst-case deviation between the empirical expectations and the true expectations:

$$\text{minimize} \quad \mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_{\bullet,i}, U_i) - \mathbb{E}f(X_\bullet, U) \right|$$

over all admissible encoding and decoding strategies given the rate constraint R and the information patterns at the two nodes (i.e., which node knows what). In other words, the goal is to ensure that, from the observer’s viewpoint, the empirical distribution of $\{(X_{\bullet,i}, U_i)\}_{i=1}^n$ is as close as possible to the target distribution $P_{X_\bullet U}$ in the sense that the corresponding expectations of all $f \in \mathcal{F}$ are as close as possible, *uniformly* over \mathcal{F} . Operational criteria of this kind arise, e.g., in the context of statistical learning from random samples [12], [13], where the functions in \mathcal{F} may be viewed as candidate predictors of U given X_\bullet .

In this paper, we consider two special cases of this set-up:

- 1) Given two alphabets \mathcal{X} and \mathcal{Y} , we take $\mathcal{X}_A = \mathcal{X}$, $\mathcal{X}_B = \emptyset$, $U = \mathcal{Y}$, $\bullet = A$. This is a generalization of the basic two-node empirical coordination problem [3, Section III.C] to abstract alphabets. (A related problem, though with a slightly different operational criterion, is lossy source coding with respect to a family of distortion measures [14].)
- 2) We have \mathcal{X} and \mathcal{Y} as above, but now $\mathcal{X}_A = U = \mathcal{X}$, $\mathcal{X}_B = \mathcal{Y}$, and $\bullet = B$. Moreover, $P_{U|X_A X_B} = P_{U|X_B} = P_{X_A|X_B}$ ¹. This is a generalization of the problem of communication of probability distributions [15] to abstract alphabets, where we also allow side information at the decoder (Node B).

Our achievability results hinge on the assumption that the function class \mathcal{F} admits the *Uniform Law of Large Numbers* (ULLN). Given an abstract alphabet Z , we say that a class \mathcal{F} of functions $f : Z \rightarrow [-1, 1]$ admits the ULLN if the following

holds: for any i.i.d. random process $\{Z_i\}_{i=1}^\infty$ over Z , we have

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(Z_i) - \mathbb{E}f(Z_1) \right| \xrightarrow{n \rightarrow \infty} 0, \quad \text{a.s.} \quad (1)$$

The quantity inside the $|\cdot|$ is referred to as the *empirical process* associated with Z^n , and describes the fluctuations of the sample mean of each f around its expectation. We define an n -tuple $z^n = (z_1, \dots, z_n) \in Z^n$ to be ε -typical w.r.t. \mathcal{F} for a probability law P if

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(z_i) - \mathbb{E}_P f(Z) \right| < \varepsilon. \quad (2)$$

Turning now to the set-up of Figure 1, let us assume that the observer’s function class \mathcal{F} satisfies the ULLN and that $\bullet = B$. Then a simple achievability argument exploits the fact (which we prove under mild regularity conditions) that, for any probability law $Q = Q_{X_A X_B U}$ under which $X_B \rightarrow X_A \rightarrow U$ is a Markov chain, there exists a rate- R encoding $\hat{U}^n(X_A^n)$ from \mathcal{X}_A^n into U^n such that the tuple (X_B^n, \hat{U}^n) is ε -typical w.r.t. \mathcal{F} for Q , provided $R > I(X_A; U|X_B)$. When $X_B = \emptyset$, so $\bullet = A$ (as in the empirical coordination scenario), we simply apply the above argument to “degenerate” Markov chains of the form $X_A \rightarrow X_A \rightarrow U$, where the rate condition becomes $R > I(X_A; U)$.

We list the salient features of our approach:

- When the underlying alphabet Z is finite, the ULLN is satisfied by the class of *all* functions $f : Z \rightarrow [-1, 1]$, and our definition of typicality reduces to strong typicality [2], [3].
- When Z is a complete separable metric space, the ULLN is satisfied by the class of all Lipschitz functions $f : Z \rightarrow [-1, 1]$ with $\|f\|_\infty \leq 1$ and Lipschitz constant bounded by 1. Moreover, the ULLN in this case is equivalent to almost sure weak convergence of empirical distributions (Varadarajan’s theorem [16, Theorem 11.4.1]).
- In general, there is a veritable plethora of function classes satisfying the ULLN (we present several examples in Section III-A). For instance, when $Z = \mathbb{R}^d$, the ULLN is satisfied by the indicator functions of all halfspaces, balls, or rectangles (and of finite unions or intersections thereof). One example, particularly relevant in source coding, is the collection of indicator functions of Voronoi cells induced by an arbitrary set of m points in \mathbb{R}^d , for any fixed m — indeed, any such cell is an intersection of $O(m)$ halfspaces. Hence, our results apply to the setting where $\mathcal{X}_\bullet \times U \subseteq \mathbb{R}^d$ and each $(X_{\bullet,i}, U_i)$ is observed through an m -point nearest-neighbor quantizer.

B. Related work

The focus of the present paper is exclusively on source coding. However, a recent preprint of Mitran [17] uses weak convergence to develop an extension of typical sequences to Polish alphabets and then applies that definition to several channel coding problems, including an achievability result for Gel’fand–Pinsker channels [18] with input cost constraints. What distinguishes Mitran’s work from ours is his careful

¹More precisely, we require that, under $P_{X_A X_B U}$, X_A and U are conditionally i.i.d. given X_B .

use of several equivalent characterizations of weak convergence via the portmanteau theorem [16, Theorem 11.1.1]. In particular, his approach requires an explicit construction of a countable generating set for the underlying Borel σ -algebra that consists of the continuity sets of the probability law of interest. As a consequence, he is able to establish a generalization of the Markov lemma [19], [20], which in turn allows him to use binning just like in the finite-alphabet case. By contrast, our notion of typicality is considerably broader (and, in fact, contains the one based on weak convergence as a special case), but, since we do not make any major structural assumptions beyond those needed for the ULLN, we cannot establish anything as strong as the Markov lemma. However, our proof technique does not rely on the Markov lemma in its strong form, and is more in the spirit of Wyner and Ziv [21]–[23].

We also note that a restricted notion of typicality based on weak convergence was used by Kontoyiannis and Zamir [24] in the context of universal vector quantization using entropy codes. The idea there is to consider sequences of increasing length, whose empirical distributions converge in the weak topology to the output distribution of an optimal test channel in a Shannon rate-distortion problem.

C. Contents of the paper

The remainder of the paper is organized as follows. Section II sets up the notation and lists the preliminaries. In Section III we formally define function classes that satisfy the ULLN and give several examples. Then, in Section IV we motivate and formally describe our approach to typicality and establish a number of key properties, including a lemma on the preservation of typicality in a Markov structure. Next, in Section V, using this lemma as the main technical tool, we illustrate the power of the proposed new approach by proving three theorems concerning fundamental limits on minimal achievable rates for (i) two-node empirical coordination; (ii) rate-constrained distributed approximation of empirical processes with side information at the decoder; and (iii) lossy source coding under a family of distortion measures. Although these results apply to general (uncountably infinite) alphabets, the proofs are as intuitive and simple as in the finite-alphabet scenario. We follow up with some concluding remarks in Section VI. Lengthy proofs and discussions of auxiliary technical results are relegated to the Appendices.

II. PRELIMINARIES AND NOTATION

All spaces in this paper are assumed to be *standard Borel spaces* (for detailed treatments, see the lecture notes of Preston [7] or Chapter 4 of Gray [8]):

Definition 1. A measurable space (Z, \mathcal{B}_Z) is *standard Borel* if it can be metrized with a metric d such that (1) (Z, d) is a complete separable metric space, and (2) \mathcal{B}_Z coincides with the Borel σ -algebra of (Z, d) (the smallest σ -algebra containing all open sets).

Remark 1. A Polish space (i.e., a separable topological space whose topology can be metrized with a complete metric) is

automatically standard Borel. In fact, the most general known class of standard Borel spaces consists of Borel subsets of Polish spaces [8, Theorem 4.3].

From now on, when dealing with a (standard Borel) space Z , we will often not mention its Borel σ -algebra explicitly. In particular, we will tacitly assume that all probability measures on Z are defined w.r.t. \mathcal{B}_Z . The main objects associated with Z that are of interest to us are as follows:

- $\mathcal{P}(Z)$ is the space of all probability measures on Z
- $M(Z)$ is the space of all measurable functions $f : Z \rightarrow \mathbb{R}$
- $M^b(Z) \subset M(Z)$ is the normed space of all bounded measurable functions $f : Z \rightarrow \mathbb{R}$ with the sup norm

$$\|f\|_\infty \triangleq \sup_{z \in Z} |f(z)| \quad (3)$$

- $M^{b,1}(Z) \triangleq \{f \in M^b(Z) : \|f\|_\infty \leq 1\}$.

Other notation will be introduced as needed.

Standard Borel spaces possess just enough useful structure for our purposes. In particular, their σ -algebras are countably generated and contain all singletons. They also admit the existence of regular conditional distributions: If $Z = X \times Y$ with the product σ -algebra, then the probability law $P \in \mathcal{P}(Z)$ of any random couple $(X, Y) \in Z$ can be disintegrated as

$$P(A \times B) = \int_A P_{Y|X}(B|x) P_X(dx), \forall A \in \mathcal{B}_X, B \in \mathcal{B}_Y \quad (4)$$

where $P_X \in \mathcal{P}(X)$ is the marginal distribution of X and $P_{Y|X}(\cdot|x) : \mathcal{B}_Y \times X \rightarrow [0, 1]$ is a *Markov kernel*, i.e., $P_{Y|X}(\cdot|x) \in \mathcal{P}(Y)$ for all $x \in X$ and $P_{Y|X}(B|\cdot) \in M(X)$ for all $B \in \mathcal{B}_Y$. Given a random triple $(U, X, Y) \in U \times X \times Y$ with joint law $P \in \mathcal{P}(U \times X \times Y)$, we will say that they form a *Markov chain* in that order (and write $U \rightarrow X \rightarrow Y$) if

$$P_{U|XY}(A|x, y) = P_{U|X}(A|x), \quad \forall A \in \mathcal{B}_U \quad (5)$$

for P -almost all x, y .

We will often use de Finetti's linear functional notation for expectations [25, Section 1.4]. That is, for any $P \in \mathcal{P}(Z)$ and a P -integrable function $f : Z \rightarrow \mathbb{R}$,

$$P(f) \triangleq \mathbb{E}_P f(Z) \equiv \int_Z f dP, \quad (6)$$

and we will extend this notation in an obvious way to integrals with respect to signed Borel measures on Z . Given a class \mathcal{F} of measurable functions $f \in M^{b,1}(Z)$, we can define a seminorm on the space of all signed measures on Z via

$$\|\nu\|_{\mathcal{F}} \triangleq \sup_{f \in \mathcal{F}} |\nu(f)|. \quad (7)$$

As an example, $\|P - P'\|_{M^{b,1}(Z)}$ is precisely the *total variation distance*

$$\|P - P'\|_{TV} \triangleq 2 \sup_{A \in \mathcal{B}_Z} |P(A) - P'(A)| \quad (8)$$

between $P, P' \in \mathcal{P}(Z)$.

We will make use of several standard information-theoretic definitions [5]. The *divergence* between P and P' in $\mathcal{P}(Z)$ is defined as

$$D(P\|P') \triangleq \begin{cases} P(\log(dP/dP')), & \text{if } P \ll P' \\ +\infty, & \text{otherwise} \end{cases} \quad (9)$$

Given a $Q \in \mathcal{P}(X \times Y)$, the *mutual information* between $X \in X$ and $Y \in Y$ with joint law Q is

$$I(Q) \triangleq D(Q \| Q_X \otimes Q_Y), \quad (10)$$

where $Q_X \otimes Q_Y$ is the product of the marginals. Whenever Q is clear from context, we will also write $I(X; Y)$ instead of $I(Q)$. We will use standard notation for such things as the conditional mutual information.

III. UNIFORM LAWS OF LARGE NUMBERS AND GLIVENKO–CANTELLI CLASSES

Given an n -tuple $z^n = (z_1, \dots, z_n) \in Z^n$, let us denote by P_{z^n} the induced *empirical measure*:

$$P_{z^n} \triangleq \frac{1}{n} \sum_{i=1}^n \delta_{z_i}, \quad (11)$$

where δ_{z_i} is the Dirac measure concentrated at z_i (since B_Z contains all singletons, $\delta_z \in \mathcal{P}(Z)$ for every $z \in Z$). If $\{Z_i\}_{i=1}^\infty$ is an i.i.d. sequence with common distribution $P \in \mathcal{P}(Z)$, then the Strong Law of Large Numbers says that, for any $f \in M^{b,1}(Z)$, the empirical means

$$P_{Z^n}(f) = \frac{1}{n} \sum_{i=1}^n f(Z_i), \quad n \in \mathbb{N} \quad (12)$$

converge to the true mean $P(f)$ almost surely. By the union bound, this holds for any *finite* family of functions. In this paper, we consider *infinite* function classes that admit a Uniform Law of Large Numbers — that is, absolute deviations between empirical and true means converge to zero *uniformly* over the function class. The canonical example of such a class appears in the celebrated Glivenko–Cantelli theorem [16, Theorem 11.4.2]: Let Z be a real-valued random variable with CDF F_Z , and let $\{Z_i\}_{i=1}^\infty$ be an infinite sequence of i.i.d. copies of Z . For each n , consider the *empirical CDF*

$$F_{Z^n}(z) \triangleq \frac{1}{n} \sum_{i=1}^n 1_{\{Z_i \leq z\}}. \quad (13)$$

The Glivenko–Cantelli theorem then says that

$$\sup_{z \in \mathbb{R}} |F_{Z^n}(z) - F_Z(z)| \xrightarrow{n \rightarrow \infty} 0 \quad \text{a.s.} \quad (14)$$

To cast it as a statement about a function class, consider

$$\mathcal{F} \triangleq \{f_z = 1_{(-\infty, z]} : z \in \mathbb{R}\}. \quad (15)$$

Then for any $z \in \mathbb{R}$,

$$F_{Z^n}(z) = P_{Z^n}(f_z) \quad (16)$$

$$F_Z(z) = P_Z(f_z) \quad (17)$$

and consequently

$$\sup_{z \in \mathbb{R}} |F_{Z^n}(z) - F_Z(z)| = \|P_{Z^n} - P\|_{\mathcal{F}} \xrightarrow{n \rightarrow \infty} 0 \quad \text{a.s.} \quad (18)$$

This motivates the following definition [9]–[11]:

Definition 2. A class \mathcal{F} of measurable functions $f \in M^{b,1}(Z)$ is called *Glivenko–Cantelli*² (or *GC*, for short) if

$$\|P_{Z^n} - P\|_{\mathcal{F}} \xrightarrow{n \rightarrow \infty} 0 \quad \text{a.s.} \quad (19)$$

for every $P \in \mathcal{P}(Z)$, where $\{Z_i\}_{i=1}^\infty$ is an i.i.d. random process with marginal distribution P .

Remark 2. In view of this definition, the classical Glivenko–Cantelli theorem can be paraphrased as follows: The class of all indicator functions of semi-infinite intervals of the form $(-\infty, z]$, $z \in \mathbb{R}$, is GC.

Remark 3. The restriction to bounded functions is mostly needed for technical convenience and can be removed by means of suitable moment conditions and straightforward, though tedious, truncation arguments. A nice side benefit of the boundedness assumption, though, is that no loss of generality occurs if the almost sure convergence in (19) is replaced with convergence in probability [10], [26].

Remark 4. It should be borne in mind that when the function class \mathcal{F} is uncountable, $\|P_{Z^n} - P\|_{\mathcal{F}}$ may not be a random variable (there is always a risk of spawning a nonmeasurable monster whenever one dabbles in uncountable operations). There are a number of ways to deal with such issues, as detailed in [9, Appendix] or [10, Section 2.3]. For our purposes, it will suffice to assume that \mathcal{F} is countable or “nice” in the sense that it contains a countable subset \mathcal{G} such that for every $f \in \mathcal{F}$ there is a sequence $\{g_m\}$ in \mathcal{G} converging to f pointwise. Then

$$\|P_{Z^n} - P\|_{\mathcal{F}} = \|P_{Z^n} - P\|_{\mathcal{G}}, \quad (20)$$

and the r.h.s. is a measurable function of Z^n [10, p. 110].

Let $(\Omega, \mathcal{B}, \mathbb{P})$ be an underlying probability space for the random process $\{Z_i\}$. Then for each n we can construct another random process on $(\Omega, \mathcal{B}, \mathbb{P})$, indexed by \mathcal{F} :

$$\Delta_f^{(n)}(\omega) \triangleq P_{Z^n(\omega)}(f) - P(f), \quad f \in \mathcal{F}. \quad (21)$$

This is an instance of an *empirical process* [9]–[11], which is used to describe the fluctuations of the empirical means $P_{Z^n}(f)$ around the expectation $P(f)$. A GC class is one for which the $\ell^\infty(\mathcal{F})$ norms

$$\|\Delta_f^{(n)}(\omega)\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |\Delta_f^{(n)}(\omega)| \quad (22)$$

of the empirical processes $\{\Delta_f^{(n)}\}_{f \in \mathcal{F}}$, $n \geq 1$, converge to zero almost surely.

A. Examples of Glivenko–Cantelli classes

We close this section by listing several examples of GC classes. Usually, whether or not a given class \mathcal{F} is GC depends on how “large” it is. The simplest notion of size is captured by the (metric) *entropy numbers* of \mathcal{F} [27]. Given any $\varepsilon > 0$, the covering number $N(\varepsilon, \mathcal{F}, Q)$ of $\mathcal{F} \subset M^{b,1}(Z)$ w.r.t. a probability measure $Q \in \mathcal{P}(Z)$ is the minimal number of balls

²Strictly speaking, the proper term is “universal Glivenko–Cantelli,” but we will follow standard usage and just say “Glivenko–Cantelli.”

$\{g : \|g - f\|_{L^1(Q)} \leq \varepsilon\}$, $f \in M^{b,1}(Z)$, of radius ε needed to cover \mathcal{F} . The entropy number of \mathcal{F} is $\log N(\varepsilon, \mathcal{F}, Q)$. Then (under additional measurability assumptions, cf. Remark 4) \mathcal{F} is GC if

$$\sup_{Q \in \mathcal{P}(Z)} N(\varepsilon, \mathcal{F}, Q) < \infty, \quad \forall \varepsilon > 0. \quad (23)$$

Other conditions for a class to be GC involve alternative notions of entropy, such as entropy with bracketing. Chapter 2 of van der Waart and Wellner [10] contains a detailed exposition of these matters. Examples 1–4 below follow [10]; Example 5 shows that the well-known theorem of Varadarajan on almost sure weak convergence of empirical measures can be stated in the form of a ULLN for an appropriate GC class.

Example 1 (Vapnik–Chervonenkis classes). Given any collection $\mathcal{A} \subset \mathcal{B}_Z$ and any finite set $C \subset Z$, define

$$S(\mathcal{A}, C) \triangleq |\{C \cap A : A \in \mathcal{A}\}| \quad (24)$$

$$S_n(\mathcal{A}) \triangleq \max_{|C| \leq n} S(\mathcal{A}, C) \quad (25)$$

and let $V(\mathcal{A}) \triangleq \max\{n \in \mathbb{N} : S_n(\mathcal{A}) = 2^n\}$. After the fundamental work of Vapnik and Chervonenkis [28] where these combinatorial parameters were first introduced, any class \mathcal{A} such that $V(\mathcal{A}) < \infty$ is called a *Vapnik–Chervonenkis (VC) class*, and $V(\mathcal{A})$ is called its *Vapnik–Chervonenkis (VC) dimension*. Examples of VC classes include:

- The class of all rectangles in \mathbb{R}^d with VC dimension $2d$.
- The class of all linear halfspaces $H_{w,b} = \{z \in \mathbb{R}^d : \langle w, z \rangle + b \geq 0\}$ for $w \in \mathbb{R}^d$, $b \in \mathbb{R}$, with VC dimension $d + 1$.
- The class of all closed balls $B_{x,r} = \{z \in \mathbb{R}^d : \|z - x\| \leq r\}$ for $x \in \mathbb{R}^d$, $r \in \mathbb{R}^+$, with VC dimension $d + 1$.

Given a collection $\mathcal{A} \subset \mathcal{B}_Z$, let $\mathcal{F} \equiv \mathcal{F}_{\mathcal{A}}$ consist of the indicator functions of the elements of \mathcal{A} : $\mathcal{F}_{\mathcal{A}} = \{1_A : A \in \mathcal{A}\}$. Then $\mathcal{F}_{\mathcal{A}}$ is GC, provided \mathcal{A} is a VC class.

Finite set-theoretic operations (unions, intersections, complements) on VC classes yield VC classes as well. In particular, consider the collection of all Voronoi cells induced by all m -point subsets of \mathbb{R}^d . Each member of this collection is an intersection of $O(m)$ halfspaces, and therefore we have a VC class. Likewise, injective images of VC classes are VC.

Example 2 (VC-subgraph classes). Given a function $f \in M(Z)$, its *subgraph* is the subset of $Z \times \mathbb{R}$, given by $\{(z, t) : f(z) > t\}$. A class of functions $\mathcal{F} \subset M(Z)$ is called a *VC-subgraph class* if the collection of all subgraphs of all $f \in \mathcal{F}$ is a VC class in $Z \times \mathbb{R}$. We define $V(\mathcal{F})$, the VC dimension of \mathcal{F} , as the VC dimension of the corresponding collection of subgraphs. For example, if \mathcal{F} is a linear span of m functions $f_1, \dots, f_m \in M(Z)$, then it is a VC-subgraph class with $V(\mathcal{F}) \leq m + 2$. In this paper, we are interested primarily in the case when $\mathcal{F} \subset M^{b,1}(Z)$. Hence, if $f_1, \dots, f_m \in M^{b,1}(Z)$, then their convex hull is a VC-subgraph class.

Example 3 (VC-hull classes). A class of functions $\mathcal{F} \subset M(Z)$ is a *VC-hull class* if there exists a VC-subgraph class $\mathcal{G} \subset M(Z)$, such that every $f \in \mathcal{F}$ is a pointwise limit of a

sequence of functions $\{f_n\}$ contained in the *symmetric convex hull* of \mathcal{G} ,

$$\left\{ \sum_{i=1}^m c_i g_i : m \in \mathbb{N}; \sum_{i=1}^m |c_i| \leq 1; g_1, \dots, g_m \in \mathcal{G} \right\} \quad (26)$$

For example, the set of all monotone functions $f : \mathbb{R} \rightarrow [0, 1]$ is VC-hull (though not VC-subgraph).

Example 4 (Smooth functions). Let $Z = [0, 1]^d$. For any multi-index, i.e., a vector $\underline{k} = (k_1, \dots, k_d) \in \{0, 1, \dots\}^d$, define the differential operator

$$D^{\underline{k}} \triangleq \frac{\partial^{|\underline{k}|}}{\partial z_1^{k_1} \dots \partial z_d^{k_d}}, \quad (27)$$

where $|\underline{k}| \triangleq k_1 + \dots + k_d$. Given $\alpha > 0$, define for a function $f : [0, 1]^d \rightarrow \mathbb{R}$

$$\|f\|_{\alpha} \triangleq \max_{\underline{k}: |\underline{k}| \leq \lfloor \alpha \rfloor} \sup_z |D^{\underline{k}} f(z)| + \max_{\underline{k}: |\underline{k}| = \lfloor \alpha \rfloor} \sup_{z \neq z'} \frac{|D^{\underline{k}} f(z) - D^{\underline{k}} f(z')|}{\|z - z'\|^{\alpha - \lfloor \alpha \rfloor}} \quad (28)$$

Let C^{α} be the set of all continuous functions $f : [0, 1]^d \rightarrow \mathbb{R}$ with $\|f\|_{\alpha} \leq 1$. Then C^{α} is a GC class.

Example 5 (Bounded Lipschitz functions). Let (Z, d) be a complete separable metric space. Define the *Lipschitz seminorm* $\|\cdot\|_L$ on $M(Z)$ by

$$\|f\|_L \triangleq \sup_{z \neq z'} \frac{|f(z) - f(z')|}{d(z, z')} \quad (29)$$

and the *bounded Lipschitz norm* $\|\cdot\|_{BL}$ by

$$\|f\|_{BL} \triangleq \|f\|_{\infty} + \|f\|_L. \quad (30)$$

Note that any function f with $\|f\|_{BL} < \infty$ is automatically in $C^b(Z)$, the Banach space of all bounded continuous functions on Z .

Let $\mathcal{F}_{BL}^1 = \{f \in C^b(Z) : \|f\|_{BL} \leq 1\}$. Then \mathcal{F} is a GC class. This is a consequence of the fact that the *bounded Lipschitz metric* (also known as the Fortet–Mourier metric)

$$\beta(P, P') \triangleq \sup_{f \in \mathcal{F}_{BL}^1} |P(f) - P'(f)| \quad (31)$$

$$\equiv \|P - P'\|_{\mathcal{F}_{BL}^1}, \quad P, P' \in \mathcal{P}(Z) \quad (32)$$

metrizes the topology of weak convergence in $\mathcal{P}(Z)$. Recall that a sequence $\{P_n\}$ in $\mathcal{P}(Z)$ converges *weakly* to $P \in \mathcal{P}(Z)$ (the fact denoted by $P_n \rightsquigarrow P$) if

$$P_n(f) \xrightarrow{n \rightarrow \infty} P(f), \quad \forall f \in C^b(Z). \quad (33)$$

Then $P_n \rightsquigarrow P$ if and only if $\beta(P_n, P) \xrightarrow{n \rightarrow \infty} 0$ [16, Theorem 11.3.3]. Now, according to a theorem of Varadarajan [16, Theorem 11.4.1], given any i.i.d. random process $\{Z_i\}_{i=1}^{\infty}$ over Z with common marginal distribution $P \in \mathcal{P}(Z)$, the empirical distributions P_{Z^n} converge weakly to P almost surely:

$$P_{Z^n} \rightsquigarrow P \quad \text{a.s.} \quad (34)$$

From the foregoing discussion, (34) is equivalent to

$$\beta(P_{Z^n}, P) = \sup_{f \in \mathcal{F}_{\text{BL}}^1} |P_{Z^n}(f) - P(f)| \quad (35)$$

$$\equiv \|P_{Z^n} - P\|_{\mathcal{F}_{\text{BL}}^1} \xrightarrow{n \rightarrow \infty} 0 \quad \text{a.s.} \quad (36)$$

In other words, $\mathcal{F}_{\text{BL}}^1$ is a GC class, and Varadarajan's theorem can be paraphrased to say that this function class obeys a ULLN.

IV. RETHINKING TYPICALITY FOR GENERAL ALPHABETS

Now that all necessary definitions are made, we can introduce our revised notion of typicality for standard Borel spaces.

For finite alphabets, there are multiple equivalent definitions of a typical sequence. Here is one, based on the total variation distance [3], often referred to as *strong typicality* [2, Section 10.6]:

Definition 3. Given a finite set Z and a probability distribution (mass function) P on it, the typical set $\mathcal{T}_\varepsilon^{(n)}(P)$, for $\varepsilon > 0$, is the set of all n -tuples $z^n \in Z^n$ whose empirical distributions P_{z^n} are ε -close to P in total variation:

$$\mathcal{T}_\varepsilon^{(n)}(P) \triangleq \{z^n \in Z^n : \|P_{z^n} - P\|_{\text{TV}} < \varepsilon\}. \quad (37)$$

By the Law of Large Numbers, if $\{Z_i\}$ is a sequence of i.i.d. draws from P , then

$$\mathbb{P}(Z^n \notin \mathcal{T}_\varepsilon^{(n)}(P)) \xrightarrow{n \rightarrow \infty} 0. \quad (38)$$

If Z is a Cartesian product $X \times Y$, then one can define *jointly* and *conditionally* typical sets and sequences [2].

However, all of this breaks down for general (uncountably infinite) alphabets. The reason is that the total variation distance between any discrete measure and a nonatomic measure is equal to 2. Indeed, if (Z, \mathcal{B}_Z) is a standard Borel space and $P \in \mathcal{P}(Z)$ assigns zero mass to singletons, $P(\{z\}) = 0, \forall z \in Z$, then we can take any n -tuple $z^n \in Z^n$ and let A be the set of its *distinct* elements, so that $P_{z^n}(A) = 1$ and $P(A) = 0$. Using this and the definition (8), we deduce that $\|P_{z^n} - P\|_{\text{TV}} = 2$.

Of course, one could use typicality arguments by considering arbitrary finite quantizations of the underlying spaces, but, as long as we are dealing with nonatomic measures, this does not get rid of the above issue even in the limit of increasingly fine quantizations. While discretization is sufficient for many purposes [5], there is another issue that arises when dealing with Markov structures in multiterminal settings: quantization destroys the Markov property [29, Section VIII].

To resolve this conundrum, we recall (cf. Sec. II) that

$$\|P - P'\|_{\text{TV}} = \sup_{\|f\|_\infty \leq 1} |P(f) - P'(f)|, \quad (39)$$

where the supremum is over *all* measurable functions $f : Z \rightarrow [-1, 1]$. When the underlying measurable space supports nonatomic probability measures, this function class turns out to be too large to admit uniform convergence of empirical averages to statistical expectations. A natural solution, then, is to restrict the class of functions:

Definition 4. Let Z be a Borel space and let $\mathcal{F} \subset M^{b,1}(Z)$ be a GC class of functions. Given a probability measure $P \in$

$\mathcal{P}(Z)$, the typical set $\mathcal{T}_{\varepsilon, \mathcal{F}}^{(n)}(P)$, for $\varepsilon > 0$, is the set of all n -tuples $z^n \in Z^n$ whose empirical distributions P_{z^n} are ε -close to P in the $\|\cdot\|_{\mathcal{F}}$ seminorm:

$$\mathcal{T}_{\varepsilon, \mathcal{F}}^{(n)}(P) \triangleq \{z^n \in Z^n : \|P_{z^n} - P\|_{\mathcal{F}} < \varepsilon\}. \quad (40)$$

One thing to note is that when Z is finite, we can just take $\mathcal{F} = M^{b,1}(Z)$ and immediately recover Definition 3. Moreover, if Z is a complete separable metric space, then we can take $\mathcal{F} = \mathcal{F}_{\text{BL}}^1$, in which case our notion of typicality becomes compatible with the bounded Lipschitz metric that metrizes the weak topology on the space of probability laws (cf. Example 5).

A. Basic properties of GC typical sets

We now establish several basic properties of GC typical sets. First of all, any sufficiently long sequence emitted by a stationary memoryless source is typical with high probability:

Proposition 1. Consider a Borel space Z and a GC class $\mathcal{F} \subset M^{b,1}(Z)$. If $\{Z_i\}_{i=1}^\infty$ is an i.i.d. random process over Z with common law P , then for any $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} \mathbb{P}(Z^n \notin \mathcal{T}_{\varepsilon, \mathcal{F}}^{(n)}(P)) = 0 \quad (41)$$

Proof: Immediate from definitions. ■

Another desirable property is for typicality to be preserved under coordinate projections. It is not hard to show that, for any two finite alphabets X and Y and any two n -tuples $x^n \in X^n$ and $y^n \in Y^n$ that are jointly typical w.r.t. some $P \in \mathcal{P}(X \times Y)$ in the sense of Definition 3, x^n (resp., y^n) is typical w.r.t. the marginal distribution P_X (resp., P_Y). The following lemma gives a sufficient condition for GC typicality to be preserved under projections:

Proposition 2. Suppose $Z = X \times Y$. Let $\pi_X : Z \rightarrow X$ be the coordinate projection mapping onto X , i.e., $\pi_X(x, y) = x$, and extend it to tuples via

$$\pi_X((x_1, y_1), \dots, (x_n, y_n)) = (x_1, \dots, x_n). \quad (42)$$

Then for any $n \in \mathbb{N}$, any $\varepsilon > 0$, any $P \in \mathcal{P}(Z)$, and any GC class $\mathcal{F}_X \subset M^{b,1}(X)$ such that $\mathcal{F}_X \circ \pi_X \subseteq \mathcal{F}$, we have the inclusion

$$\pi_X(\mathcal{T}_{\varepsilon, \mathcal{F}}^{(n)}(P)) \subseteq \mathcal{T}_{\varepsilon, \mathcal{F}_X}^{(n)}(P_X). \quad (43)$$

Remark 5. As can be seen from the proof below, the class \mathcal{F}_X need not be GC in order for the inclusion (43) to hold. However, then one would not be able to transfer a convergence result like Proposition 1 to the X -valued part of the sequence.

Proof: Suppose $z^n = ((x_1, y_1), \dots, (x_n, y_n)) \in$

$\mathcal{T}_{\varepsilon, \mathcal{F}}^{(n)}(P)$. Then

$$\|P_{x^n} - P_X\|_{\mathcal{F}_X} = \sup_{f \in \mathcal{F}_X} \left| \frac{1}{n} \sum_{i=1}^n f(x_i) - P_X(f) \right| \quad (44)$$

$$= \sup_{f \in \mathcal{F}_X} \left| \frac{1}{n} \sum_{i=1}^n f \circ \pi_X(z_i) - P(f \circ \pi_X) \right| \quad (45)$$

$$\leq \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(z_i) - P(f) \right| \quad (46)$$

$$= \|P_{z^n} - P\|_{\mathcal{F}} \quad (47)$$

$$< \varepsilon. \quad (48)$$

Thus, $x^n \in \mathcal{T}_{\varepsilon, \mathcal{F}_X}^{(n)}(P_X)$, which proves (43). ■

As an example, let $X = \mathbb{R}^k$, $Y = \mathbb{R}^m$, let \mathcal{F} be the collection of indicator functions of all halfspaces in $Z = \mathbb{R}^{k+m}$, and let \mathcal{F}_X be the collection of indicator functions of all halfspaces in X (cf. Example 1 for definitions and notation). For any $w \in \mathbb{R}^k$, $b \in \mathbb{R}$ and $z = (x, y) \in Z$, we have

$$\langle w, x \rangle + b = \langle w, \pi_X(z) \rangle + b \quad (49)$$

$$= \langle (w, 0), (x, y) \rangle + (b, 0). \quad (50)$$

Hence, $1_{H_{(w,0),(b,0)}} = 1_{H_{(w,b)}} \circ \pi_X$ for any choice of $w \in \mathbb{R}^k$, $b \in \mathbb{R}$, so the condition of the lemma is satisfied.

Finally, we show that our definition of typicality can work in a multiterminal setting. Ideally, one would like to have something like the Markov lemma [19], [20]: If $X \rightarrow Y \rightarrow Z$ is a Markov chain, (x^n, y^n) is typical, and Z^n is obtained by passing y^n through a memoryless channel, then (x^n, y^n, Z^n) should be typical with high probability. However, in our setting such a statement does not make much sense without assuming additional structure for the function class \mathcal{F} .³ Instead, we establish the following result, which is essentially an abstract alphabet version of the so-called Piggyback Coding Lemma of Wyner [21, Lemma 4.3]:

Lemma 1. *Let $U \in \mathcal{U}$, $V \in \mathcal{V}$, and $W \in \mathcal{W}$ be random variables taking values in their respective standard Borel spaces according to a joint distribution P_{UVW} , such that $U \rightarrow V \rightarrow W$ is a Markov chain and $I(V; W) < \infty$. Let $\{(U_i, V_i, W_i)\}_{i=1}^\infty$ be a sequence of i.i.d. draws from P_{UVW} . Let $\mathcal{F} \subset M^{b,1}(\mathcal{U} \times \mathcal{W})$ be a GC class of functions. For a given $\varepsilon > 0$, there exist an $n = n(\varepsilon)$ and a mapping $\Phi_n : \mathcal{V}^n \rightarrow \mathcal{W}^n$, such that*

$$\frac{1}{n} \log |\{\Phi_n(v^n) : v^n \in \mathcal{V}^n\}| \leq I(V; W) + \varepsilon \quad (51)$$

and

$$\mathbb{P}\left((U^n, \Phi_n(V^n)) \notin \mathcal{T}_{\varepsilon, \mathcal{F}}^{(n)}(P_{UW})\right) < \varepsilon. \quad (52)$$

Proof: For each n , define the function $\psi_n \in M^{b,1}(\mathcal{U}^n \times \mathcal{W}^n)$ by

$$\psi_n(u^n, w^n) \triangleq 1_{\{(u^n, w^n) \notin \mathcal{T}_{\varepsilon, \mathcal{F}}^{(n)}(P_{UW})\}}. \quad (53)$$

³Incidentally, this is exactly what Mitran [17] accomplishes for his notion of typicality based on weak convergence.

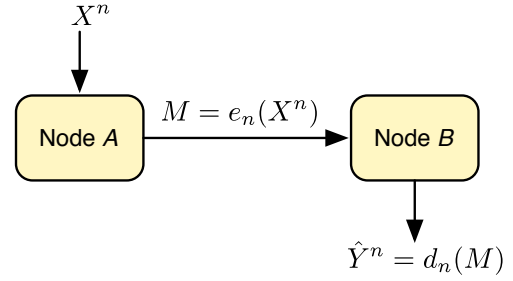


Fig. 2. Two-node empirical coordination.

Since \mathcal{F} is a GC class, we have by Proposition 1

$$\lim_{n \rightarrow \infty} \mathbb{E} \psi_n(U^n, W^n) = 0. \quad (54)$$

The desired statement now follows from Lemma A.1 in Appendix A. ■

V. APPLICATIONS TO EMPIRICAL COORDINATION

We now show three sample applications of GC typicality to the problem of empirical coordination in a two-node network shown in Figure 1. This problem, recently formulated and studied by Cuff et al. [3], concerns joint generation of actions at the two nodes, such that the empirical distribution of the actions over time approximates, asymptotically, a desired joint distribution in total variation. Our goal is to extend this setting to general alphabets. As we have shown in Section IV, the total variation criterion is unsuitable for uncountable alphabets, so we consider a relaxation to an appropriate GC class.

As we will show, our notion of GC typicality and Lemma 1 can be used to develop particularly intuitive achievability arguments and to obtain single-letter characterizations of the best achievable rates. Moreover, convexity of the $\|\cdot\|_{\mathcal{F}}$ seminorm is helpful for proving converse results. The downside, however, is that, in general, it is not possible to compute the best achievable rates explicitly even for “simple” sources due to the presence of the supremum over \mathcal{F} .

A. Two-node empirical coordination

Consider the two-node network shown in Fig. 2, where Node A (resp., Node B) generates actions from a Borel space X (resp., Y). At Node A, the actions are drawn i.i.d. from a fixed law $P_X \in \mathcal{P}(X)$. We also have a conditional probability measure $P_{Y|X}$ that describes the desired distribution of actions at Node B given the actions at Node A. Following the terminology of [3], we will also refer to the choice of $P_{Y|X}$ as a *coordination*. Node A can communicate with Node B over a rate-limited channel, and Node B uses the data it receives to choose its actions. For each n , let $X^n \in X^n$ and $Y^n \in Y^n$ denote the action sequences at the two nodes. Given a class $\mathcal{F} \subset M^{b,1}(X \times Y)$ of measurable “test functions” and a desired distortion level $\Delta \geq 0$, the goal is for Node A to communicate with Node B at a minimal rate to guarantee that, asymptotically,

$$\mathbb{E} \|P_{(X^n, Y^n)} - P_X \otimes P_{Y|X}\|_{\mathcal{F}} \lesssim \Delta, \quad (55)$$

where $P_{XY} = P_X \otimes P_{Y|X}$ is the joint law induced by the source P_X and the coordination $P_{Y|X}$.

Definition 5. An (n, M) -code is a pair (e_n, d_n) , where $e_n : \mathcal{X}^n \rightarrow [M]$ is the encoder and $d_n : [M] \rightarrow \mathcal{Y}^n$ is the decoder, and $[M] \triangleq \{1, 2, \dots, M\}$. We will denote $\hat{Y}^n = d_n(e_n(X^n))$.

Definition 6. Given a source P_X , a coordination $P_{Y|X}$, and a distortion Δ , let $\mathcal{E}(\Delta, P_{Y|X})$ denote the set of all $Q \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, such that

$$Q_X = P_X \quad \text{and} \quad \|Q - P_X \otimes P_{Y|X}\|_{\mathcal{F}} \leq \Delta. \quad (56)$$

Define the rate-distortion/coordination function of P_X :

$$R(\Delta, P_{Y|X}) \triangleq \inf_{Q \in \mathcal{E}(\Delta, P_{Y|X})} I(Q). \quad (57)$$

Theorem 1. Let $P_{Y|X}$ be a given coordination and Δ a given distortion level.

- a) **Direct part:** If \mathcal{F} is a GC class and $R(\Delta, P_{Y|X}) < \infty$, then for any $\varepsilon > 0$ there exist $n \equiv n(\varepsilon)$ and an $(n, 2^{nR})$ -code (e_n, d_n) with $R < R(\Delta, P_{Y|X}) + \varepsilon$ satisfying

$$\mathbb{E}\|P_{(X^n, \hat{Y}^n)} - P_X \otimes P_{Y|X}\|_{\mathcal{F}} \leq \Delta + \varepsilon. \quad (58)$$

- b) **Converse part:** Suppose that there exists an $(n, 2^{nR})$ -code $\hat{Y}^n(X^n) = d_n(e_n(X^n))$, satisfying

$$\mathbb{E}\|P_{(X^n, \hat{Y}^n)} - P_X \otimes P_{Y|X}\|_{\mathcal{F}} \leq \Delta. \quad (59)$$

Then $R \geq R(\Delta, P_{Y|X})$.

Remark 6. Note that the converse does not require \mathcal{F} to be GC. However, it must be sufficiently “well-behaved” for $\|P_{(X^n, \hat{Y}^n)} - P_X \otimes P_{Y|X}\|_{\mathcal{F}}$ to be measurable for any choice of a (measurable) encoder-decoder pair.

Proof (direct part): To prove the direct part, fix $(\Delta, P_{Y|X})$ and pick any $Q \in \mathcal{E}(\Delta, P_{Y|X})$ such that $I(Q) < R(\Delta, P_{Y|X}) + \varepsilon/2$. Let $X \in \mathcal{X}$ and $U \in \mathcal{Y}$ have joint law Q . Then $X \rightarrow X \rightarrow U$ is a Markov chain, and Lemma 1 guarantees the existence of an n and a mapping $\Phi_n : \mathcal{X}^n \rightarrow \mathcal{Y}^n$, such that

$$\frac{1}{n} \log |\{\Phi_n(X^n)\}| \leq I(Q) + \varepsilon/2 \quad (60)$$

$$< R(\Delta, P_{Y|X}) + \varepsilon \quad (61)$$

and

$$\mathbb{E}\|P_{(X^n, \Phi_n(X^n))} - Q\|_{\mathcal{F}} \leq \varepsilon. \quad (62)$$

Let $\hat{Y}^n = \Phi_n(X^n)$. Then the triangle inequality gives

$$\begin{aligned} \mathbb{E}\|P_{(X^n, \hat{Y}^n)} - P_X \otimes P_{Y|X}\|_{\mathcal{F}} \\ \leq \mathbb{E}\|P_{(X^n, \hat{Y}^n)} - Q\|_{\mathcal{F}} + \|Q - P_X \otimes P_{Y|X}\|_{\mathcal{F}} \end{aligned} \quad (63)$$

$$\leq \Delta + \varepsilon, \quad (64)$$

which establishes (58). ■

Proof (converse part): For the converse, we will use the time mixing technique (cf. [3] and Appendix B). Let $\hat{Y}^n(X^n)$ be an $(n, 2^{nR})$ -code such that (59) holds. Let T be a random variable uniformly distributed over the set $[n]$, independently

of X^n , and let \hat{Q} denote the joint distribution of (X_T, \hat{Y}_T) . Then

$$nR \geq H(\hat{Y}^n(X^n)) \quad (65)$$

$$= H(\hat{Y}^n(X^n)) - H(\hat{Y}^n(X^n)|X^n) \quad (66)$$

$$= I(X^n; \hat{Y}^n(X^n)) \quad (67)$$

$$\geq \sum_{t=1}^n I(X_t; \hat{Y}_t) \quad (68)$$

$$= nI(X_T; \hat{Y}_T|T) \quad (69)$$

$$= nI(X_T; \hat{Y}_T, T) \quad (70)$$

$$\geq nI(X_T; \hat{Y}_T) \quad (71)$$

$$= nI(\hat{Q}), \quad (72)$$

where:

- (65) holds because the log-cardinality of the range of $\hat{Y}^n(\cdot)$ is bounded by nR
- (68) is a standard information-theoretic fact: if X^n is an i.i.d. tuple, then for any sequence $\hat{Y}_1, \dots, \hat{Y}_n$ jointly distributed with X^n

$$I(X^n; \hat{Y}^n) \geq \sum_{t=1}^n I(X_t; \hat{Y}_t) \quad (73)$$

- (69) follows from the construction of T
- (70) holds because, by the chain rule for mutual information,

$$I(X_T; \hat{Y}_T, T) = I(X_T; T) + I(X_T; \hat{Y}_T|T),$$

where the first term on the r.h.s. is zero because X^n is i.i.d. (see Fact 1 in Appendix B).

The remaining steps are consequences of other definitions and standard information-theoretic identities.

Since X^n is i.i.d., X_T is independent of T and has the same distribution as X_1 , namely P_X . Moreover, the expected empirical distribution $\mathbb{E}P_{(X^n, \hat{Y}^n)}$ is equal to $P_{(X_T, \hat{Y}_T)} \equiv \hat{Q}$ (Fact 2 in Appendix B). Thus, we can write

$$\begin{aligned} \|\hat{Q} - P_X \otimes P_{Y|X}\|_{\mathcal{F}} \\ = \|\mathbb{E}P_{(X^n, \hat{Y}^n)} - P_X \otimes P_{Y|X}\|_{\mathcal{F}} \end{aligned} \quad (74)$$

$$\leq \mathbb{E}\|P_{(X^n, \hat{Y}^n)} - P_X \otimes P_{Y|X}\|_{\mathcal{F}} \quad (75)$$

$$\leq \Delta, \quad (76)$$

where (75) follows from convexity, and (76) from (59). Hence, $\hat{Q} \in \mathcal{E}(\Delta, P_{Y|X})$, so $R \geq I(\hat{Q}) \geq R(\Delta, P_{Y|X})$. ■

B. Communication of empirical processes

The next application we consider also concerns distributed approximation of an empirical process. We have a joint law $P = P_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. Let $\{(X_i, Y_i)\}_{i=1}^\infty$ be an infinite sequence of independent draws from P . Consider the two-node network shown in Figure 3. Node A (resp., Node B) has perfect observations of $\{X_i\}$ (resp., $\{Y_i\}$). As before, Node A can transmit information to Node B over a rate-limited channel. The goal is for Node A to communicate with Node B at a minimal rate, so that Node B can approximate the desired empirical process to within a given distortion level Δ .

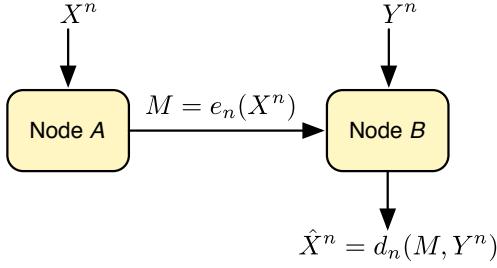


Fig. 3. Distributed compression of empirical processes.

More precisely, given a block length n and denoting by \hat{X}^n the reconstruction of X^n at Node B, we wish to guarantee that

$$\mathbb{E} \|\mathbb{P}_{(\hat{X}^n, Y^n)} - P\|_{\mathcal{F}} \lesssim \Delta. \quad (77)$$

This setting is a generalization of the problem of *communication of probability distributions*, recently formulated and studied by Kramer and Savari [15] in the finite-alphabet setting. Here, we allow general alphabets and decoder side information. As we will see, the minimum achievable rate admits a single-letter characterization reminiscent of the Wyner–Ziv rate-distortion function for lossy source coding with decoder side information [22], [23].

Definition 7. An (n, M) -code is a pair (e_n, d_n) , where $e_n : X^n \rightarrow [M]$ is the encoder and $d_n : [M] \times Y^n \rightarrow X^n$ is the decoder. We will denote $\hat{X}^n = d_n(e_n(X^n), Y^n)$.

Definition 8. Given a source $P_{XY} \in \mathcal{P}(X \times Y)$, let $\mathcal{E}(\Delta)$ denote the set

$$\{Q \in \mathcal{P}(X \times Y \times U) : U \text{ is standard Borel}\}$$

such that:

- 1) $Q_{XY} = P_{XY}$
- 2) $Q_{U|XY} = Q_{U|X}$ (i.e., $Y \rightarrow X \rightarrow U$ is a Markov chain)
- 3) There is a function $g : Y \times U \rightarrow X$, such that

$$\|Q_{WY} - P\|_{\mathcal{F}} \leq \Delta, \quad (78)$$

where $W = g(Y, U)$.

With this, define the rate-distortion function

$$R(\Delta) \triangleq \inf_{Q \in \mathcal{E}(\Delta)} [I(Q_{XU}) - I(Q_{YU})]. \quad (79)$$

Theorem 2. Let \mathcal{F} be a class of functions $f : X \times Y \rightarrow [0, 1]$ and Δ a nonnegative distortion level.

- a) **Direct part:** Suppose that \mathcal{F} is a GC class, and that for any $\delta > 0, \mu \in \mathcal{P}(X \times Y)$ one can find a finite set $\{\hat{x}_j\}_{j=1}^N \subset X$ and a quantizer $q : X \rightarrow \{\hat{x}_j\}$, such that

$$\|\mu_{q(X)Y} - \mu\|_{\mathcal{F}} \leq \delta. \quad (80)$$

If $R(\Delta) < \infty$, then for any $\varepsilon > 0$ there exist an $n \equiv n(\varepsilon)$ and an $(n, 2^{nR})$ code with $R < R(\Delta) + \varepsilon$ satisfying

$$\mathbb{E} \|\mathbb{P}_{(\hat{X}^n, Y^n)} - P\|_{\mathcal{F}} \leq \Delta + \varepsilon, \quad (81)$$

where $\hat{X}^n = d_n(e_n(X^n), Y^n)$.

- b) **Converse part:** Suppose that there exists an $(n, 2^{nR})$ -code $\hat{X}^n = d_n(e_n(X^n), Y^n)$ satisfying

$$\mathbb{E} \|\mathbb{P}_{(\hat{X}^n, Y^n)} - P\|_{\mathcal{F}} \leq \Delta. \quad (82)$$

Then $R \geq R(\Delta)$.

Remark 7. The quantization assumption (80) is a “smoothness” condition on \mathcal{F} , and is akin to an assumption made by Wyner in [23] in order to extend the achievability part of the finite-alphabet result of [22] to abstract alphabets.

Proof (direct part): First we show that, owing to the quantization assumption (80), we can assume w.l.o.g. that both Y and the auxiliary alphabet U are finite. This follows from the following lemma, whose proof is given in Appendix C:

Lemma 2. Consider any law $Q \in \mathcal{E}(\Delta)$. Then, for any $\delta > 0$, there exist finite measurable partitions $\{A_i\}_{i=1}^{N_1}$ and $\{B_j\}_{j=1}^{N_2}$ of Y and U and a function $g_1 : Y \times U \rightarrow X$ such that:

- a) $\|Q_{W_1Y} - P\|_{\mathcal{F}} \leq \Delta + \delta$, where $W_1 = g_1(Y, U)$
- b) g_1 is constant on the rectangles $A_i \times B_j$, $1 \leq i \leq N_1, 1 \leq j \leq N_2$
- c) $I(Q_{X\tilde{U}}) - I(Q_{Y\tilde{U}}) \leq I(Q_{XU}) - I(Q_{YU}) + \delta$ where $\tilde{Y} = i$ for $Y \in A_i$ and $\tilde{U} = j$ for $U \in B_j$.

Let us therefore assume that U and Y are both finite. We will use a Wyner–Ziv style two-step argument [22], [23]: The first step consists of using a long block code that preserves typicality (following Lemma 1), while the second step uses a Slepian–Wolf code [30] to communicate the codewords with negligible probability of error. Pick any $Q \in \mathcal{E}(\Delta)$ such that

$$I(Q_{XU}) - I(Q_{YU}) < R(\Delta) + \varepsilon/2. \quad (83)$$

Define $\bar{g} : Y \times U \rightarrow X \times Y$ by $\bar{g}(y, u) \triangleq (g(y, u), y)$ and consider the function class $\mathcal{F} \circ \bar{g} \subset M^{b,1}(Y \times U)$. Since \mathcal{F} is a GC class, so is $\mathcal{F} \circ \bar{g}$ — to see this, fix any $\mu \in \mathcal{P}(Y \times U)$ and let $\{(Y_i, U_i)\}_{i=1}^\infty$ be a sequence of i.i.d. draws from μ . Then for any n we can write

$$\begin{aligned} & \|\mathbb{P}_{(Y^n, U^n)} - \mu\|_{\mathcal{F} \circ \bar{g}} \\ &= \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(g(Y_i, U_i), Y_i) - \mathbb{E} f(g(Y, U), Y) \right| \end{aligned} \quad (84)$$

$$= \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(W_i, Y_i) - \mathbb{E} f(W, Y) \right| \quad (85)$$

$$\equiv \|\mathbb{P}_{(W^n, Y^n)} - \mu_{WY}\|_{\mathcal{F}}, \quad (86)$$

where $W = g(Y, U)$. Thus, the GC property of $\mathcal{F} \circ \bar{g}$ follows from the GC property of \mathcal{F} .⁴ In view of this, we can apply Lemma 1 to the Markov chain $Y \rightarrow X \rightarrow U$ and to the GC class $\mathcal{F} \circ \bar{g}$ to derive the existence of a large enough n_1 and a mapping $\Phi_{n_1} : X^{n_1} \rightarrow U^{n_1}$, such that

$$\frac{1}{n_1} \log |\{\Phi_{n_1}(X^{n_1})\}| \leq I(Q_{XU}) + \varepsilon/2 \quad (87)$$

⁴By contrast, in order for the GC property to be preserved under left compositions, i.e., for $\psi \circ \mathcal{F}$ to be a GC class for some $\psi : [0, 1] \rightarrow [0, 1]$, additional requirements must be imposed on ψ (such as monotonicity or Lipschitz continuity).

and

$$\begin{aligned} \mathbb{E} \|P_{(Y^{n_1}, \hat{U}^{n_1})} - Q_{YU}\|_{\mathcal{F} \circ \bar{g}} \\ = \mathbb{E} \|P_{(\hat{W}^{n_1}, Y^{n_1})} - Q_{WY}\|_{\mathcal{F}} \leq \varepsilon/2, \end{aligned} \quad (88)$$

where

$$\hat{W}^{n_1} = (g(Y_1, \hat{U}_1), \dots, g(Y_{n_1}, \hat{U}_{n_1})) \quad (89)$$

$$\hat{U}^{n_1} = \Phi_{n_1}(X^{n_1}). \quad (90)$$

We can use a blocking argument along the lines of Lemmas 3 and 5 of Wyner and Ziv [22] to show that a sufficiently long sequence $\hat{U}^{n_1}(1), \dots, \hat{U}^{n_1}(n_2)$ of i.i.d. realizations of \hat{U}^n can be losslessly encoded, using a Slepian–Wolf code, at a rate of

$$\frac{1}{n_1} H(\hat{U}^{n_1} | Y^{n_1}) \leq I(Q_{XU}) - I(Q_{YU}) + \varepsilon/2 \quad (91)$$

$$< R(\Delta) + \varepsilon. \quad (92)$$

Let $n = n_1 n_2$, and let $\{\tilde{U}_i\}_{i=1}^n$ denote the resulting decoding. Then, if n_2 is large enough, we can guarantee that

$$\mathbb{E} \|P_{(Y^n, \tilde{U}^n)} - P_{(Y^n, \hat{U}^n)}\|_{\mathcal{F} \circ \bar{g}} \leq \varepsilon/2, \quad (93)$$

and therefore, with $\hat{X}^n = (g(Y_1, \tilde{U}_1), \dots, g(Y_n, \tilde{U}_n))$, that

$$\begin{aligned} \mathbb{E} \|P_{(\hat{X}^n, Y^n)} - Q_{WY}\|_{\mathcal{F}} \\ = \mathbb{E} \|P_{(Y^n, \tilde{U}^n)} - Q_{YU}\|_{\mathcal{F} \circ \bar{g}} \end{aligned} \quad (94)$$

$$\begin{aligned} \leq \mathbb{E} \|P_{(Y^n, \tilde{U}^n)} - P_{(Y^n, \hat{U}^n)}\|_{\mathcal{F} \circ \bar{g}} \\ + \mathbb{E} \|P_{(Y^n, \hat{U}^n)} - Q_{YU}\|_{\mathcal{F} \circ \bar{g}} \end{aligned} \quad (95)$$

$$\leq \varepsilon. \quad (96)$$

The triangle inequality then yields

$$\begin{aligned} \mathbb{E} \|P_{(\hat{X}^n, Y^n)} - P\|_{\mathcal{F}} \\ \leq \mathbb{E} \|P_{(\hat{X}^n, Y^n)} - Q_{WY}\|_{\mathcal{F}} + \|Q_{WY} - P\|_{\mathcal{F}} \end{aligned} \quad (97)$$

$$\leq \Delta + \varepsilon. \quad (98)$$

This gives a $(n, 2^{nR})$ -code with $R < R(\Delta) + \varepsilon$. ■

Proof (converse part): To prove the converse, we again use time mixing. Let (e_n, d_n) be an $(n, 2^{nR})$ code, let $J = e_n(X^n)$ and $\hat{X}^n = d_n(J, Y^n)$, and let T be uniformly distributed on $[n]$ independently of (X^n, Y^n) . Define an auxiliary random variable

$$U = (J, X^{T-1}, Y^{T-1}, Y_{T+1}^n, T) \quad (99)$$

(cf. [3], [22], [23]) and note that $Y_T \rightarrow X_T \rightarrow U$ is a Markov chain. Moreover,

$$nR \geq H(J) \quad (100)$$

$$\geq H(J|Y^n) \quad (101)$$

$$= I(X^n; J|Y^n) \quad (102)$$

$$= \sum_{t=1}^n I(X_t; J|Y^n, X^{t-1}) \quad (103)$$

$$= \sum_{t=1}^n I(X_t; J, X^{t-1}, Y^{t-1}, Y_{t+1}^n | Y_t) \quad (104)$$

$$= nI(X_T; J, X^{T-1}, Y^{T-1}, Y_{T+1}^n | Y_T, T) \quad (105)$$

$$= nI(X_T; J, X^{T-1}, Y^{T-1}, Y_{T+1}^n, T | Y_T) \quad (106)$$

$$= nI(X_T; U | Y_T), \quad (107)$$

where:

- (100) holds because the log-cardinality of the range of $e_n(\cdot)$ is bounded by nR
- (104) follows from the chain rule and the fact that $X_t \rightarrow Y_t \rightarrow (X^{t-1}, Y^{t-1}, Y_{t+1}^n)$ is a Markov chain
- (105) follows from the construction of T
- (106) follows because, by the chain rule,

$$\begin{aligned} I(X_T; J, X^{T-1}, Y^{T-1}, Y_{T+1}^n, T | Y_T) \\ = I(X_T; T | Y_T) + I(X_T; J, X^{T-1}, Y^{T-1}, Y_{T+1}^n | Y_T, T), \end{aligned}$$

where the first term on the r.h.s. is zero because $(X_1, Y_1), \dots, (X_n, Y_n)$ are i.i.d., so (X_T, Y_T) is independent of T (see Fact 1 in Appendix B).

The remaining steps are consequences of other definitions and standard information-theoretic identities.

Since $\{(X_i, Y_i)\}_{i=1}^n$ are i.i.d., (X_T, Y_T) has the same joint law as (X_1, Y_1) , namely P_{XY} . Moreover, \hat{X}_T is a deterministic function of (Y_T, U) , and $\mathbb{E} P_{(\hat{X}^n, Y^n)} = P_{(\hat{X}_T, Y_T)}$. Finally,

$$\|P_{(\hat{X}_T, Y_T)} - P\|_{\mathcal{F}} = \|\mathbb{E} P_{(\hat{X}^n, Y^n)} - P\|_{\mathcal{F}} \quad (108)$$

$$\leq \mathbb{E} \|P_{(\hat{X}^n, Y^n)} - P\|_{\mathcal{F}} \quad (109)$$

$$\leq \Delta, \quad (110)$$

where (109) follows from convexity, and (110) follows from (82). Hence, the joint law of X_T , Y_T , and U belongs to $\mathcal{E}(\Delta)$, which means that $R \geq I(X_T; U | Y_T) \geq R(\Delta)$. ■

C. Lossy coding with respect to a class of distortion measures

Finally, we consider the problem of lossy coding with respect to a class of distortion measures (fidelity criteria). For general (Polish) alphabets, it was solved by Dembo and Weissman [14], but the finite-alphabet variant appears already as Problem 14 in [31]. Let X and Y denote the source and the reproduction alphabets, respectively. Suppose a class Γ of distortion measures $\rho : X \times Y \rightarrow [0, 1]$ is given, together with a class of nonnegative reals indexed by $\rho \in \Gamma$, $\{\Delta_\rho\}_{\rho \in \Gamma}$. The goal is to find a block code of minimal rate whose expected distortion under each $\rho \in \Gamma$ is bounded by the corresponding Δ_ρ . We use the same definition of an (n, M) -code as in Section V-A.

Define a mapping $F(\cdot, \{\Delta_\rho\}) : \mathcal{P}(X \times Y) \rightarrow \mathbb{R}$ by

$$F(Q, \{\Delta_\rho\}) \triangleq \sup_{\rho \in \Gamma} [Q(\rho) - \Delta_\rho], \quad (111)$$

where

$$Q(\rho) = \int \rho dQ = \int \rho(x, y) Q(dx, dy) \quad (112)$$

is the expected distortion between X and Y when they have joint law Q .

Definition 9. Given a source $P_X \in \mathcal{P}(X)$, let $\mathcal{E}(\{\Delta_\rho\})$ denote the set of all $Q \in \mathcal{P}(X \times Y)$ such that

$$Q_X = P_X \quad \text{and} \quad F(Q, \{\Delta_\rho\}) \leq 0. \quad (113)$$

Define the rate-distortion function

$$R(\{\Delta_\rho\}) \triangleq \inf_{Q \in \mathcal{E}(\{\Delta_\rho\})} I(Q). \quad (114)$$

Theorem 1 of [14] shows that any rate $R \geq R(\{\Delta_\rho\})$ is achievable, provided the mapping $Q \mapsto F(Q, \{\Delta_\rho\})$ is upper semicontinuous (u.s.c.) under the weak topology on $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$. Moreover, no rate $R < R(\{\Delta_\rho\})$ is achievable. We now show that the u.s.c. requirement can be replaced by a GC condition:

Theorem 3. *Let Γ be a class of distortion measures and $\{\Delta_\rho\}_{\rho \in \Gamma}$ a class of nonnegative distortion levels.*

- a) **Direct part:** *If Γ is a GC class and $R(\{\Delta_\rho\}) < \infty$, then for any $\varepsilon > 0$, there exist an $n \equiv n(\varepsilon)$ and an $(n, 2^{nR})$ code with $R < R(\{\Delta_\rho\}) + \varepsilon$ satisfying*

$$\mathbb{E} \sup_{\rho \in \Gamma} [\rho(X^n, \hat{Y}^n) - \Delta_\rho] \leq \varepsilon, \quad (115)$$

where $\rho(X^n, \hat{Y}^n) \triangleq P_{(X^n, \hat{Y}^n)}(\rho)$.

- b) **Converse part:** *Suppose that there exists an $(n, 2^{nR})$ -code $\hat{Y}^n = d_n(e_n(X^n))$ satisfying*

$$\mathbb{E} \rho(X^n, \hat{Y}^n) \leq \Delta_\rho, \quad \forall \rho \in \Gamma. \quad (116)$$

Then $R \geq R(\{\Delta_\rho\})$.

Proof: To prove the direct part, pick any $Q \in \mathcal{E}(\{\Delta_\rho\})$ such that $I(Q) < R(\{\Delta_\rho\}) + \varepsilon/2$. Let $X \in \mathcal{X}$ and $U \in \mathcal{Y}$ have joint law Q . The same argument as in the proof of Theorem 1 can be used to show the existence of a large enough n and a mapping $\Phi_n : \mathcal{X}^n \rightarrow \mathcal{Y}^n$, such that

$$\frac{1}{n} \log |\{\Phi_n(X^n)\}| \leq I(Q) + \varepsilon/2 \quad (117)$$

$$< R(\{\Delta_\rho\}) + \varepsilon \quad (118)$$

and

$$\mathbb{E} \|\mathbb{P}_{(X^n, \hat{Y}^n)} - Q\|_\Gamma \leq \varepsilon, \quad (119)$$

where $\hat{Y}^n = \Phi_n(X^n)$. Now, for any $\rho \in \Gamma$ we have

$$\rho(X^n, \hat{Y}^n) - \Delta_\rho \leq \|\mathbb{P}_{(X^n, \hat{Y}^n)} - Q\|_\Gamma + F(Q, \{\Delta_\rho\}). \quad (120)$$

Consequently, taking the supremum of both sides over Γ and then the expectation w.r.t. P_{X^n} , we get (115).

The proof of the converse is exactly the same as in [14]. ■

VI. CONCLUSION

We have proposed a new definition of typical sequences over a wide class of abstract alphabets (standard Borel spaces), which retains many useful properties of strong (total-variation) typicality for finite alphabets. In particular, it is preserved in a Markov structure, which has allowed us to develop transparent achievability proofs in several settings pertaining to empirical coordination of actions in a two-node network using finite communication resources. Here are some directions for future research:

- *Behavior in the finite block length regime* — GC classes with sufficiently “regular” metric or combinatorial structure admit sharp concentration-of-measure inequalities of the form

$$\mathbb{P}(\|\mathbb{P}_{Z^n} - P\|_{\mathcal{F}} \geq \varepsilon) \leq S(n; \mathcal{F}) e^{-Cn\varepsilon^2}, \quad (121)$$

where $C > 0$ is some constant and $S(n; \mathcal{F})$ is a function of “moderate” growth in n , which typically depends on the geometric characteristics of \mathcal{F} [9]–[11]. For example, if \mathcal{F} is a VC class, then $S(n; \mathcal{F}) = O(n^{V(\mathcal{F})})$; in the latter case, we also have

$$\mathbb{E} \|\mathbb{P}_{Z^n} - P\|_{\mathcal{F}} \leq C \sqrt{\frac{V(\mathcal{F})}{n}}, \quad (122)$$

where $C > 0$ is a universal constant. These inequalities can be used to investigate the behavior of our coding schemes in the finite block length regime (e.g., the rate of convergence of the achievable $\|\cdot\|_{\mathcal{F}}$ -distortion to the optimum).

- *Extension to stationary ergodic sources* — Recently, Adams and Nobel [32] have shown that the ULLN holds for countable (or separable) classes of VC sets and functions even when the underlying process is stationary and ergodic (rather than i.i.d.), although without any specific guarantees on the rate of convergence. Their work opens the possibility of extending our GC typicality approach to stationary ergodic sources via sliding block codes [33]–[35].
- *Connections to simulation of information sources* — The operational criteria used in our treatment of empirical coordination suggest new ways of thinking about simulation of random processes and related problems in rate-distortion coding [3], [36]–[38]. Many problems related to sensing, learning, and control under communication constraints can be reduced (or related) to simulation of random processes, and our formalism may be of use for characterizing the fundamental information-theoretic limits in these settings.

APPENDIX A

PIGGYBACK CODING LEMMA FOR BOREL SPACES

In this appendix we prove the following lemma, which is an extension of the Piggyback Coding lemma of Wyner [21, Lemma 4.3] to general alphabets:

Lemma A.1. *Let $\mathcal{U}, \mathcal{V}, \mathcal{W}$ be standard Borel spaces, and let $(U, V, W) \in \mathcal{U} \times \mathcal{V} \times \mathcal{W}$ be a triple of random variables with joint law P_{UVW} , such that $U \rightarrow V \rightarrow W$ is a Markov chain and the mutual information $I(V; W)$ is finite. Let $\{(U_i, V_i, W_i)\}_{i=1}^\infty$ be a sequence of i.i.d. draws from P_{UVW} . Let $\{\psi_n\}_{n=1}^\infty$ be a sequence of measurable functions $\psi_n : \mathcal{U}^n \times \mathcal{W}^n \rightarrow [0, 1]$, such that*

$$\lim_{n \rightarrow \infty} \mathbb{E} \psi_n(U^n, W^n) = 0. \quad (A.1)$$

For a given $\varepsilon > 0$, there exists $n_0 = n_0(\varepsilon)$, such that for every $n \geq n_0$ we can find a mapping $F_n : \mathcal{V}^n \rightarrow \mathcal{W}^n$ that satisfies

$$\frac{1}{n} \log \left| \left\{ F_n(v^n) : v^n \in \mathcal{V}^n \right\} \right| \leq I(V; W) + \varepsilon \quad (A.2)$$

and

$$\mathbb{E} \psi_n(U^n, F_n(V^n)) \leq \varepsilon. \quad (A.3)$$

Proof: The proof is very similar to Wyner's proof for finite alphabets [21]. Fix any n and define a function $\phi_n : \mathbb{V}^n \times \mathbb{W}^n \rightarrow [0, 1]$ by

$$\begin{aligned} \phi_n(v^n, w^n) &\triangleq \mathbb{E}[\psi_n(U^n, W^n) | V^n = v^n, W^n = w^n] \quad (\text{A.4}) \\ &= \int_{\mathbb{U}^n} \psi_n(u^n, w^n) P_{U^n | V^n, W^n}(du^n | v^n, w^n). \end{aligned} \quad (\text{A.5})$$

Owing to the Markov chain condition, we can write

$$\phi_n(v^n, w^n) = \int_{\mathbb{U}^n} \psi_n(u^n, w^n) P_{U^n | V^n}(du^n | v^n). \quad (\text{A.6})$$

Letting $\delta_n \triangleq \mathbb{E}\psi_n(U^n, W^n)$, we define the set

$$\mathcal{S}_n \triangleq \{v^n, w^n \in \mathbb{V}^n \times \mathbb{W}^n : \phi_n(v^n, w^n) \leq \sqrt{\delta_n}\}. \quad (\text{A.7})$$

Then by the Markov inequality we have

$$\mathbb{P}((V^n, W^n) \notin \mathcal{S}_n) \leq \frac{\mathbb{E}\phi_n(V^n, W^n)}{\sqrt{\delta_n}} = \sqrt{\delta_n}. \quad (\text{A.8})$$

Consider an arbitrary measurable mapping $G : \mathbb{V}^n \rightarrow \{w^n(1), \dots, w^n(M)\} \subset \mathbb{W}^n$ for some $M < \infty$. Then, defining the set

$$\tilde{\mathcal{S}}_n \triangleq \{v^n \in \mathbb{V}^n : (v^n, G(v^n)) \in \mathcal{S}_n\}, \quad (\text{A.9})$$

we can write

$$\mathbb{E}\psi_n(U^n, G(V^n)) \quad (\text{A.10})$$

$$= \mathbb{E}[\mathbb{E}[\psi_n(U^n, G(V^n)) | V^n]] \quad (\text{A.11})$$

$$= \mathbb{E}\phi_n(V^n, G(V^n)) \quad (\text{A.12})$$

$$\leq \mathbb{P}(\tilde{\mathcal{S}}_n^c) + \int_{\tilde{\mathcal{S}}_n} \phi_n(v^n, G(v^n)) P_{V^n}(dv^n), \quad (\text{A.13})$$

where (A.12) is due to (A.6), while (A.13) uses the fact that $0 \leq \phi_n(\cdot, \cdot) \leq 1$. Moreover,

$$\begin{aligned} &\int_{\tilde{\mathcal{S}}_n} \phi_n(v^n, G(v^n)) P_{V^n}(dv^n) \\ &= \sum_{m=1}^M \int_{\tilde{\mathcal{S}}_n \cap G^{-1}(w^n(m))} \phi_n(v^n, w^n(m)) P_{V^n}(dv^n) \quad (\text{A.14}) \\ &\leq \sqrt{\delta_n}. \end{aligned} \quad (\text{A.15})$$

Hence,

$$\mathbb{E}\psi_n(U^n, G(V^n)) \leq \mathbb{P}(\tilde{\mathcal{S}}_n^c) + \sqrt{\delta_n}. \quad (\text{A.16})$$

Now we can use Lemma 9.3.1 in [39] to show that, given \mathcal{S}_n , M , and an arbitrary $R > 0$, there exist a set $\{w^n(1), \dots, w^n(M)\} \subset \mathbb{W}^n$ and a mapping $G_n : \mathbb{V}^n \rightarrow \{w^n(1), \dots, w^n(M)\}$, such that

$$\begin{aligned} &\mathbb{P}((V^n, G_n(V^n)) \notin \mathcal{S}_n) \leq \mathbb{P}(\tilde{\mathcal{S}}_n^c) \\ &\quad + \mathbb{P}(i(V^n, W^n) > nR) + \exp(-M2^{-Rn}), \end{aligned} \quad (\text{A.17})$$

where

$$i(v^n, w^n) \triangleq \log \frac{dP_{V^n, W^n}}{d(P_{V^n} \otimes P_{W^n})}(v^n, w^n) \quad (\text{A.18})$$

is the information density [5]. Letting $M = 2^{n(I(V; W) + \varepsilon)}$ and $R = I(V; W) + \varepsilon/2$ and using the corresponding mapping G_n , we get

$$\begin{aligned} \mathbb{E}\psi_n(U^n, G_n(V^n)) &\leq 2\sqrt{\delta_n} \\ &\quad + \exp(-2^{n\varepsilon/2}) + \mathbb{P}(i(V^n, W^n) > nR). \end{aligned} \quad (\text{A.19})$$

Since $\mathbb{E}\psi_n(U^n, W^n) = \delta_n \rightarrow 0$ as $n \rightarrow \infty$, the first term goes to zero as $n \rightarrow \infty$. The second term likewise goes to 0 since $\varepsilon > 0$. The third term goes to zero owing to the mean ergodic theorem for information densities [5, Theorem 8.5.1]. Choosing n_0 large enough so that the right-hand side of the above inequality is less than ε finishes the proof. ■

APPENDIX B TIME MIXING

Our discussion of the time mixing technique essentially follows [3, p. 4200], except that care must be taken due to the fact that we are working with general alphabets here.

Fix a space \mathbb{U} . Let $U^n = (U_1, \dots, U_n)$ be a random n -tuple taking values in \mathbb{U}^n according to some law P_{U^n} . Let T be a random variable uniformly distributed over the set $[n]$ independently of U^n . Consider the random variable $U_T \in \mathbb{U}$, i.e., the value of the T th coordinate of U^n . We will use two facts pertaining to this construction.

First, we note that U_T and T need not be independent, even though U^n and T are. One exception is when U^n is an i.i.d. tuple:

Fact 1. *If U^n is an i.i.d. tuple with common marginal P_U , then U_T is independent of T and has the same law as U_1 , i.e., P_U .*

Proof: For any $i \in [n]$ and any $A \in \mathcal{B}_{\mathbb{U}}$,

$$P_{U_T, T}(A \times \{i\}) = \mathbb{P}(T = i) P_{U_T | T}(A | i) \quad (\text{B.1})$$

$$= \mathbb{P}(T = i) P_{U_i}(A) \quad (\text{B.2})$$

$$= \mathbb{P}(T = i) P_U(A) \quad (\text{B.3})$$

$$= P_T(\{i\}) P_U(A). \quad (\text{B.4})$$

Hence, $P_{U_T | T}(A | i) = P_U(A)$, regardless of i . ■

Second, let us consider the empirical distribution P_{U^n} . Since \mathbb{U} is a Borel space, $\mathcal{P}(\mathbb{U})$ is a (complete separable) metric space under any metric that metrizes the weak convergence of probability laws, so we can equip it with its Borel σ -algebra. Then P_{U^n} is a $\mathcal{P}(\mathbb{U})$ -valued random variable, whose expectation $\mathbb{E}P_{U^n}$ is given by

$$[\mathbb{E}P_{U^n}](A) \triangleq \frac{1}{n} \sum_{i=1}^n P_{U_i}(A), \quad \forall A \in \mathcal{B}_{\mathbb{U}}. \quad (\text{B.5})$$

It is not hard to check that $\mathbb{E}P_{U^n}$ satisfies the Kolmogorov axioms and is itself an element of $\mathcal{P}(\mathbb{U})$. In particular:

Fact 2. *Consider the empirical distribution P_{U^n} . Then*

$$\mathbb{E}P_{U^n} = P_{U_T}, \quad (\text{B.6})$$

where $P_{U_T} \in \mathcal{P}(\mathbb{U})$ is the law of U_T .

Proof: For any $A \in \mathcal{B}_U$,

$$[\mathbb{E}P_{U^n}](A) = \frac{1}{n} \sum_{i=1}^n P_{U_i}(A) \quad (\text{B.7})$$

$$= \mathbb{E} \left[\sum_{i=1}^n \mathbb{P}(T=i) 1_{\{U_i \in A\}} \right] \quad (\text{B.8})$$

$$= \mathbb{E} [\mathbb{E}[1_{\{U_T \in A\}} | U^n]] \quad (\text{B.9})$$

$$= \mathbb{E} [1_{\{U_T \in A\}}] \quad (\text{B.10})$$

$$= P_{U_T}(A). \quad (\text{B.11})$$

Since A is arbitrary, (B.6) indeed holds. \blacksquare

APPENDIX C PROOF OF LEMMA 2

The proof is very similar to the proof of Lemma 5.3 of Wyner [23]. In particular, only part (a) requires modification. Parts (b) and (c) follow immediately, just as in [23].

Since $Q \in \mathcal{E}(\Delta)$, there exists a function $g : Y \times U \rightarrow X$, such that, with $W = g(Y, U)$,

$$\|Q_{WY} - P\|_{\mathcal{F}} \leq \Delta. \quad (\text{C.1})$$

Secondly, owing to the smoothness assumption (80), for any $\delta_1 > 0$ one can find a quantizer $q : X \rightarrow \{\hat{x}_j\}_{j=1}^N \subset X$, $N < \infty$, such that

$$\|Q_{q(W)Y} - Q_{WY}\| \leq \delta_1. \quad (\text{C.2})$$

Let $g_0 \triangleq q \circ g$, and define the sets

$$C_j \triangleq \{(y, u) \in Y \times U : g_0(y, u) = \hat{x}_j\}, \quad 1 \leq j \leq N. \quad (\text{C.3})$$

Lemma 5.4 in [23] can be used to show that, for an arbitrary $\delta_2 > 0$, there exists a collection of disjoint sets $\{S_j\}_{j=1}^N \subset \mathcal{B}_Y \otimes \mathcal{B}_U$, where each S_j is a finite union of rectangles, and

$$Q_{YU}(S_j \triangle C_j) \leq \delta_2, \quad 1 \leq j \leq N. \quad (\text{C.4})$$

Now define $g_1 : Y \times U \rightarrow X$ by

$$g_1(y, u) \triangleq \begin{cases} \hat{x}_j, & \text{if } (y, u) \in S_j \\ \hat{x}_1, & \text{if } (y, u) \notin \bigcup_{j=1}^N S_j. \end{cases} \quad (\text{C.5})$$

Define also the set $E \triangleq \bigcup_{j=1}^N (C_j \cap S_j)$ and note that $g_1 = g_0$ on E . Then

$$\begin{aligned} & \mathbb{E}[f(g_1(Y, U), Y)] \\ &= \mathbb{E}[1_E f(g_0(Y, U), Y)] + \mathbb{E}[1_{E^c} f(g_1(Y, U), Y)] \end{aligned} \quad (\text{C.6})$$

$$\leq \mathbb{E}[f(g_0(Y, U), Y)] + Q_{YU}(E^c) \quad (\text{C.7})$$

$$= \mathbb{E}[f(q(W), Y)] + Q_{YU}(E^c) \quad (\text{C.8})$$

$$\leq \mathbb{E}[f(W, Y)] + \delta_1 + Q_{YU}(E^c). \quad (\text{C.9})$$

Similarly,

$$\begin{aligned} & \mathbb{E}[f(W, Y)] \\ & \leq \mathbb{E}[f(q(W), Y)] + \delta_1 \end{aligned} \quad (\text{C.10})$$

$$= \mathbb{E}[1_E f(q(W), Y)] + \mathbb{E}[1_{E^c} f(q(W), Y)] + \delta_1 \quad (\text{C.11})$$

$$= \mathbb{E}[1_E f(g_1(Y, U), Y)] + \mathbb{E}[1_{E^c} f(q(W), Y)] + \delta_1 \quad (\text{C.12})$$

$$\leq \mathbb{E}[f(g_1(Y, U), Y)] + Q_{YU}(E^c) + \delta_1. \quad (\text{C.13})$$

In both cases we have used the fact that f is bounded between 0 and 1, as well as (C.2). Moreover, using the fact that $\{C_j\}$ is a disjoint partition of $Y \times U$, as well as (C.4), we can write

$$Q_{YU}(E^c) \leq \sum_{j=1}^J Q_{YU}(S_j \triangle C_j) \leq N\delta_2. \quad (\text{C.14})$$

Combining (C.1), (C.9) and (C.13), we get

$$\|Q_{W_1Y} - Q_{WY}\|_{\mathcal{F}} \leq \delta_1 + N\delta_2, \quad (\text{C.15})$$

where $W_1 = g_1(Y, U)$. Now, given $\delta > 0$, first choose $\delta_1 = \delta/2$. This fixes $N = N(\delta)$. Then choose δ_2 so that $N\delta_2 \leq \delta/2$. This proves part (a); parts (b) and (c) follow exactly as in [23].

ACKNOWLEDGMENT

The author would like to thank Todd Coleman and Serdar Yüksel for their careful reading of the manuscript and for making a number of useful suggestions that have improved the presentation.

REFERENCES

- [1] C. E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.
- [2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York: Wiley, 2006.
- [3] P. W. Cuff, H. H. Permuter, and T. M. Cover, "Coordination capacity," *IEEE Trans. Inform. Theory*, vol. 56, no. 9, pp. 4181–4206, September 2010.
- [4] A. R. Barron, "The strong ergodic theorem for densities: generalized Shannon–McMillan–Breiman theorem," *Ann. Probab.*, vol. 13, no. 4, pp. 1292–1303, 1985.
- [5] R. M. Gray, *Entropy and Information Theory*. New York: Springer-Verlag, 1990.
- [6] A. Dembo and O. Zeitouni, *Large Deviations: Techniques and Applications*. New York: Springer, 1998.
- [7] C. Preston, "Some notes on standard Borel and related spaces," September 2008, [Online]. Available: <http://arxiv.org/abs/0809.3066>
- [8] R. M. Gray, *Probability, Random Processes, and Ergodic Properties*, 2nd ed. Springer, 2009.
- [9] D. Pollard, *Convergence of Stochastic Processes*. New York: Springer, 1984.
- [10] A. W. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes*. New York: Springer-Verlag, 1996.
- [11] S. van de Geer, *Empirical Processes in M-Estimation*. Cambridge Univ. Press, 2000.
- [12] K. L. Buescher and P. R. Kumar, "Learning by canonical smooth estimation – Part I: simultaneous estimation," *IEEE Trans. Automat. Control*, vol. 41, no. 4, pp. 545–556, April 1996.
- [13] M. Raginsky, "Achievability results for learning under communication constraints," in *Proc. Inform. Theory and Applications Workshop*, San Diego, CA, 2009, pp. 272–279.
- [14] A. Dembo and T. Weissman, "The minimax distortion redundancy in noisy source coding," *IEEE Trans. Inform. Theory*, vol. 49, no. 11, pp. 3020–3030, November 2003.
- [15] G. Kramer and S. A. Savari, "Communicating probability distributions," *IEEE Trans. Inform. Theory*, vol. 53, no. 2, pp. 518–525, February 2007.
- [16] R. M. Dudley, *Real Analysis and Probability*, 2nd ed. Cambridge Univ. Press, 2002.
- [17] P. Mitran, "Typical sequences for Polish alphabets," *IEEE Trans. Inform. Theory*, 2009, submitted. [Online]. Available: <http://arxiv.org/abs/1005.2321>
- [18] S. I. Gel'fand and M. S. Pinsker, "Coding for channel with random parameters," *Probl. Contr. Inform. Theory*, vol. 9, no. 1, pp. 19–31, 1980.
- [19] T. Berger, "Multiterminal source coding," in *The Information Theory Approach to Communications*, G. Longo, Ed. New York: Springer, 1978.
- [20] G. Kramer, "Topics in multi-user information theory," *Foundations and Trends in Communications and Information Theory*, vol. 4, no. 4-5, pp. 265–444, 2007.

- [21] A. D. Wyner, "On source coding with side information at the decoder," *IEEE Trans. Inform. Theory*, vol. IT-21, no. 3, pp. 294–300, May 1975.
- [22] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inform. Theory*, vol. IT-22, no. 1, pp. 1–10, January 1976.
- [23] A. D. Wyner, "The rate-distortion function for source coding with side information at the decoder II: general sources," *Inform. Control*, vol. 38, pp. 60–80, 1978.
- [24] I. Kontoyiannis and R. Zamir, "Mismatched codebooks and the role of entropy coding in lossy data compression," *IEEE Trans. Inform. Theory*, vol. 52, no. 5, pp. 1922–1938, May 2006.
- [25] D. Pollard, *A User's Guide To Measure Theoretic Probability*. Cambridge Univ. Press, 2003.
- [26] J. M. Steele, "Empirical discrepancies and subadditive processes," *Ann. Probab.*, vol. 6, no. 1, pp. 118–127, 1978.
- [27] A. N. Kolmogorov and V. M. Tihomirov, " ϵ -entropy and ϵ -capacity of sets in function spaces," in *Amer. Math. Soc. Transl.*, ser. 2, 1961, vol. 17, pp. 277–364.
- [28] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory Probab. Appl.*, vol. 16, pp. 264–280, 1971.
- [29] I. Csiszár, "The method of types," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2505–2523, June 1998.
- [30] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, vol. IT-19, no. 4, pp. 471–480, July 1973.
- [31] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Sources*. Budapest: Akadémiai Kiadó, 1981.
- [32] T. M. Adams and A. B. Nobel, "Uniform convergence of Vapnik–Chervonenkis classes under ergodic sampling," *Ann. Probab.*, vol. 38, no. 4, pp. 1345–1367, 2010.
- [33] J. G. Dunham, "Abstract alphabet sliding-block entropy compression coding with a fidelity criterion," *Ann. Probab.*, vol. 8, no. 6, pp. 1085–1092, 1980.
- [34] J. C. Kieffer, "Extension of source coding theorems for block codes to sliding block codes," *IEEE Trans. Inform. Theory*, vol. IT-26, no. 6, pp. 679–692, November 1980.
- [35] —, "A method for proving multiterminal source coding theorems," *IEEE Trans. Inform. Theory*, vol. IT-27, no. 5, pp. 565–570, September 1981.
- [36] T. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Trans. Inform. Theory*, vol. 39, no. 3, pp. 752–772, March 1993.
- [37] Y. Steinberg and S. Verdú, "Simulation of random processes and rate-distortion theory," *IEEE Trans. Inform. Theory*, vol. 42, no. 1, pp. 63–86, January 1996.
- [38] M. Z. Mao, R. M. Gray, and T. Linder, "Rate-constrained simulation and source coding IID sources," 2010, submitted. [Online]. Available: <http://arxiv.org/abs/1008.2008>
- [39] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.